Audio Engineering Society

Conference Paper 1

Presented at the 6th International Conference on Audio for Games
2024 April 27–29, Tokyo, Japan

This paper was peer-reviewed as a complete manuscript for presentation at this conference. This paper is available in the AES E-Library (http://www.aes.org/e-lib), all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Investigating the Influence of Environmental Acoustics and Playback Device for Audio Augmented Reality Applications

Jacob Bhattacharyya[1], Lorenzo Picinali[2], Alessandro Vinciarelli[1], and Stephen Brewster[1]

[1]*University of Glasgow*
[2]*Imperial College London*

Correspondence should be addressed to Jacob Bhattacharyya (j.bhattacharyya.1@research.gla.ac.uk)

ABSTRACT

Presenting plausible virtual sounds to a user is an important challenge within audio augmented reality (AAR), where virtual sounds must appear as a real part of the audio environment. Reproducing an environment's acoustics is one step towards this, however there is limited understanding of how the spatial resolution and spectral bandwidth of such reproductions contribute to plausibility, and therefore which approaches an AAR developer should target. We present two studies comparing room impulse responses (varying in spatial resolution and spectral bandwidth) and playback devices (headphones and audio glasses) to investigate their influence on the plausibility and user perception of virtual sounds. We do so using both a listening test in a controlled environment, and then an AAR game played in two real-world locations. Our results suggest that, particularly in a real-world AAR application context, users have low sensitivity for differences between reverberation models, but that the reproduction of an environment's acoustics positively influences the plausibility and externalisation of a virtual sound. These benefits are most pronounced when played over headphones, but users were positive about the use of audio glasses for an AAR application, despite their lower perceptual fidelity. Overall, our findings suggest both lower fidelity environmental acoustics and audio glasses are appropriate for future AAR applications, allowing developers to use less computing resources and maintain real-world awareness without compromising user experience.

1 Introduction

Presenting virtual sounds as a plausible component of the surrounding environment is crucial to creating rich and engaging auditory experiences, particularly for audio augmented reality (AAR) applications. For AAR, where virtual sounds are introduced into a user's real-world surroundings, accurately simulating the effect of those surroundings on sound is an important step towards plausible presentation, yet one that is often overlooked within existing research. Yang *et al.'s* 2022 review of the field found that the majority of AAR applications provided minimal information on their environmental acoustic reproductions, or did not reproduce the effect of a user's surroundings on virtual sounds at all [1]. As a seamless blending of real and virtual sounds is core to AAR, and existing research shows that reproducing environmental acoustics can improve plausibility [2], clarifying the influence of environmental acoustics in an AAR context is an important step towards achieving that. Understanding how different

approaches to such reproductions influence a user's perception will allow AAR developers to make informed decisions about incorporating a user's surroundings into the experience, and maximise perceptual benefits, computational cost, and resource usage.

While virtual sounds must be a plausible component of an AAR user's surroundings, they must also be presented to a user without occluding those real-world surroundings [1, 3]. Audio glasses are one potential solution, resembling a standard pair of glasses with small speakers embedded in each leg, providing an audio experience similar to headphones without occluding the ear. While such glasses present a promising platform for AAR experiences [3], they are under-researched, and it is unclear how they affect user experience and perception compared to traditional platforms, both in critical listening and real-world use.

In this work, we present two studies investigating how spatial resolution and spectral bandwidth affect the perception of different room impulse responses (RIRs), and how this is affected by space, application context, and playback device. By determining how these factors affect the plausibility and perception of a virtual sound, we provide guidance for application developers to make informed evaluations of different acoustic reproduction procedures. The first study consisted of a localisation test and perceptual evaluation of RIRs varying in spatial resolution (omnidirectional, stereo, 1st-order Ambisonics and 3rd-order Ambisonics) and spectral bandwidth (sinusoidal sweep and handclap), conducted in a controlled indoor environment. The second study consisted of an AAR game scenario using those RIRs in two real-world public environments, where participants were presented with listening tests between game-play rounds. Both studies were presented over headphones and audio glasses, and modelled direct sound separately at the highest quality available, evaluating the influence of RIR characteristics on reflected sound only.

1.1 Research Questions

The work presented here sought to answer the following research questions:

RQ1: When reproducing environmental acoustics, how does the variation of spatial resolution and spectral bandwidth affect the plausibility and perception of virtual sounds in AAR scenarios?

RQ2: How do audio glasses compare to traditional headphones perceptually and in AAR applications?

2 Background and Related Work

While a clear definition of audio augmented reality is yet to be agreed upon, AAR is usually considered a subcategory of augmented reality (AR) applications where virtual sounds are inserted into a user's surroundings, and sound is used as the primary display modality [4]. These virtual sounds are usually rendered to the user with binaural spatialisation[1], where the acoustic influence of a listener's head is simulated for virtual sounds using a head-related transfer function (HRTF), for presentation over headphones.

As the field has developed, there have been a handful of examples of AAR games, including early work like Guided By Voices [6], re-imaginings of classic games like PacMan [7], or novel experiences capable of movement tracking and gestural control [8]. While AAR games have received research interest, there has yet to be a comprehensive examination of how environmental acoustics contribute to the experience. Paterson *et al.* found that the addition of reverberation improved players' immersion and emotional engagement, but only tested this with generic reverberation rather than reproducing the acoustics of the play space [9].

The acoustics of that play space can be described through a room impulse response (RIR), a measure of how a sound in one position evolves in a space over time, as perceived by a listener in another position. By convolving this RIR with a piece of audio, the resulting audio file will appear as if it were sounding in that space. RIRs can be measured directly or simulated using tools like CATT-Acoustic[2] and Google Resonance[3], which take a 3D model of an environment and simulate its acoustics. While RIR measurement simply requires recording an acoustically excited room and deconvolving that recording with the excitation source [10], RIRs are only valid for their given combination of source and listener position, necessitating additional measurements for other source and listener positions. Simulation of RIRs requires geometric models of the space to be created, which while often available for games and VR applications, are infeasible for general purpose AAR currently.

However it is done, the accurate reproduction of an environment's acoustics can improve a listener's sensation of externalisation (where a virtual sound appears

[1] See [5] for an overview of binaural sound.
[2] https://www.catt.se/
[3] https://resonance-audio.github.io/resonance-audio/

to be outside the head [11]), as well as improve the subjective realism of a sound [2]. This level of realism is usually defined through two terms: *plausibility* and *authenticity*. Authenticity is the perceptual identity of real sound and a virtual simulation – when presented side-by-side, a listener cannot detect any difference between the two [12]. Plausibility is the perceptual identity of a virtual simulation and the listener's expectation of the real equivalent [2] – the listener believes a virtual sound is the same as a real equivalent would be, even if a real equivalent would actually sound different. As we are focused on the user experience of an AAR system, and virtual sounds are unlikely to be presented directly alongside a real counterpart in such systems, we focus on plausibility.

Specifically, we focus on the effect of an RIR's spatial resolution and spectral bandwidth on plausibility and listener perception, something which prior work has explored. Ahrens and Andersson [13] found that differences between Ambisonic[4] order become imperceptible above 8th-order renderings, while Enge *et al.* [15] found no significant difference in plausibility in a VR context between simulated 3rd- and 7th-order RIRs, and Engel *et al.* [16] found resolutions above 3rd-order Ambisonics to show no significant degradation in perceived 'quality' compared to higher-order renderings. Compared to prior work, the studies presented here focus on *in situ* plausibility comparisons (judging stimuli compared to reality rather than only higher-order stimuli), evaluation of plausibility in controlled and real-world environments, and resulting considerations for AAR scenarios. This work also represents the first time plausibility has been evaluated in an AAR game context, and one of the first times audio glasses have been examined perceptually.

3 Methods

Two studies were carried out to investigate user perception of different measured RIRs and playback devices. The first consisted of a listening test in an indoor office environment, and the second consisted of an AAR game experience, played in two real-world outdoor locations.

Measured RIRs were chosen over simulated ones to isolate the influence of spatial resolution or spectral bandwidth, while still providing insights that could aid AAR developers when choosing a workflow for simulated RIRs. As higher resolution reproductions are more computationally demanding, perceptual benefits must be determined, and carefully traded off against computing resources, particularly in mobile or battery-constrained AAR scenarios.

3.1 Experimental Parameters

Both studies used RIR and playback device as independent variables. The second study also used play space as an independent variable, using outdoor spaces with both low and high reverberance, chosen to cover potential AAR environments. These spaces were also public, to reflect 'in-the-wild' AAR usage. RIRs varied in their spatial resolution and spectral bandwidth to directly compare the impact of impulsive source, compare the perception of 1st- and 3rd-order Ambisonics, and explore two RIRs which are less computationally demanding to render (an omnidirectional handclap and a stereo handclap). Study 2 also utilised two 'echoic' conditions, where anechoic source audio files had synthetic reverb added before convolution with RIRs, to simulate field or foley recordings which are often used by sound designers. By comparing the echoic and anechoic counterpart, we explored how the echoicity of source audio influences plausibility, and whether existing game audio and sound design workflows can be applied to AAR scenarios. Previous work shows minimal improvements when reverberation is rendered above 3rd-order Ambisonics [15, 16], and so this was chosen as the highest spatial resolution for this work.

In both studies, two playback devices were used: Sennheiser HD650 headphones and a development pair of FAUNA audio glasses[5] (hardwired to remove the influence of wireless latency). These were chosen to cover an acoustically transparent scenario (glasses), and a semi-transparent scenario (HD650s), either of which could potentially be used for AAR applications.

3.2 Audio Generation and Presentation

RIRs were captured in all three test spaces using a Zylia ZM-1[6] microphone capable of recording 3rd-order Ambisonics. A hand clap and a sinusoidal sweep (20Hz to 20kHz over 5s) were used as impulsive sources from approximately 2m away at a 0° azimuth.

[4] Ambisonics is an approach to rendering soundfields with a high spatial resolution using 4 or more microphone channels. See [14] for a more thorough overview of the technology.

[5] https://wearfauna.com

[6] https://www.zylia.co/zylia-zm-1-microphone.html

These RIRs were then truncated to cover the combinations of spatial resolution and spectral bandwidth detailed in Table 1. Omnidirectional RIRs used the W channel of the original 3rd-order recording, and stereo RIRs used the signals of two opposing microphone capsules on the ZM-1. A control or 'dry' condition without acoustic reproduction was also used. The playback level of each RIR was balanced subjectively by the researchers to keep the blend of direct and reverberant sound consistent across all conditions. For the echoic conditions, a synthetic reverb (IEM FDNReverb[7], with Room Size of 20, Reverberation Time of 1s, and Dry/Wet mix of 0.5) was applied to the anechoic source audio files. Each RIR had its direct sound component replaced with silence, so that direct sound could be reproduced separately and at the highest quality, so as to be consistent between conditions. Rather than measuring RIRs for each azimuth position a sound was presented at, the same RIR was used for all positions, effectively rotating the reverberant space around the user's head similarly to [16], where this simplification was not found to adversely impact the user experience.

In each study, virtual sounds were spatialised binaurally using the 3DTI Toolkit [17] within the Unity game engine, and the KEMAR dummy head HRTF from the SONICOM dataset [18]. Direct sound was rendered using a single instance of the 3DTI Toolkit Unity wrapper, while reverberation was decoded to 20 spatialised virtual loudspeakers in a dodecahedral layout. In both studies, participants were fitted with a Supperware headtracker[8], which tracked head movements for localisation trials, and maintained the real-world position of virtual sounds.

Both playback devices had output volume levels balanced subjectively by the researchers in order to present sounds at an equivalent loudness, and a high-pass filter at 250Hz was applied to the headphones to better match the frequency response range of the glasses, as quoted on the manufacturer website. All source audio used in these studies was anechoic to be free of any reverberation that might colour the results.

4 Study 1: Controlled Listening Test

The first study (n = 20[9]) evaluated six acoustic conditions over headphones and audio glasses. Participants were seated in a controlled indoor office environment (RT = 450ms), directly opposite a loudspeaker which played a reference track of the test stimuli at the beginning of each condition, to provide a real-world source comparison for judging plausibility. A within-subjects design was used, with participants experiencing all acoustic conditions and playback devices.

RIR Code	Excitation Source	Spatial Resolution
Omni-HC	Handclap	Omni
Omni-HC-Echoic	*Handclap*	*Omni*
Stereo-HC	Handclap	Stereo
1O-Sine	Sine Sweep	1st Order Ambisonics
3O-HC	Handclap	3rd Order Ambisonics
3O-Sine	Sine Sweep	3rd Order Ambisonics
3O-Sine-Echoic	*Sine Sweep*	*3rd Order Ambisonics*

Table 1: Details of the RIRs used for the two studies. Italicised RIRs were only used in Study 2.

For each acoustic condition, one of three sound stimuli was presented four times to the participant, who was asked to localise the sound by turning to face the perceived location and pressing a button on a computer keyboard. Sound presentation positions were balanced evenly by presenting each sound once in one of four 90° quadrants around the listener. After the participant had localised the four presentations of the sound, they were asked to rate the sound's externalisation, plausibility, realism, and their confidence in their localisations on continuous scales from 0 to 1. The process was then repeated for the remaining two sounds, then started over for the next acoustic condition, with acoustic conditions being presented in a random order. Playback device was counterbalanced, with half of participants using headphones first, and half using glasses first.

The three sound stimuli were a sample of human speech, a sample of acoustic guitar music [10], and a synthesised 'user interface' (UI) sound, reminiscent of notification sounds in existing games and applications.

4.1 Results

Localisation great-circle error [20] and questionnaire answers were analysed using three-way ANOVA tests (analysing RIR, playback device, and stimulus sound), with *post hoc* analysis conducted using Tukey HSD tests. Results are shown in Table 3 .

[7] https://plugins.iem.at/

[8] https://supperware.co.uk/headtracker-overview

[9] 12 men, 8 women, with 1 very unfamiliar, 8 unfamiliar, 4 neutral, 6 familiar, and 1 very familiar with spatial audio

[10] Both the speech sample and guitar sample are provided with the 3DTI toolkit.

Measure	Question
Externalisation	Did these sounds appear to be inside or outside your head?
Plausibility	Do you think these sounds were recorded in this room? (based on [19].)
Sound Realism	To what extent did these sounds appear to be part of the real world?
Localisation Confidence	How confident are you that you located the sounds accurately?

Table 2: Questions used in Study 1.

Measure		ANOVA p	Significant Pairwise Comparisons		
				p	Mean Diff.
Plausibility	RIR	.01	3O-HC - Omni-HC	.03	0.10
			1O-Sine - Omni-HC	.04	0.09
			3O-Sine - Omni-HC	.04	0.10
	Device	.15			
	Stimulus	< .01	Speech - Music	< .01	-0.07
Loc. Error	RIR	< .01	Stereo-HC - Dry	< .01	8.1
			1O-Sine - Dry	< .01	8.6
			3O-Sine - Dry	< .01	8.9
			3O-Sine - Omni-HC	.03	5.54
			3O-Sine - 3O-HC	.03	5.47
	Device	< .01	Glasses - HD650	< .01	6.33
	Stimulus	.05	Speech - UI	.04	-3.18
Externalisation	RIR	< .01	Omni-HC - Dry	< .01	0.11
			3O-HC - Dry	< .01	0.14
			1O-Sine - Dry	< .01	0.15
			3O-Sine - Dry	< .01	0.13
	Device	< .01	Glasses - HD650	< .01	-0.05
	Stimulus	< .01	Music - UI	< .01	0.08
			Speech - UI	< .01	0.09
Loc. Confidence	RIR	< .01	Stereo-HC - Dry	< .01	-0.18
			3O-HC - Dry	.03	-0.08
			1O-Sine - Dry	< .01	-0.13
			3O-Sine - Dry	< .01	-0.16
			Stereo-HC - Omni-HC	< .01	-0.15
			1O-Sine - Omni-HC	< .01	-0.11
			3O-Sine - Omni-HC	< .01	-0.13
			3O-HC - Stereo-HC	< 0.01	0.09
	Device	< .01	Glasses - HD650	< .01	-0.05
	Stimulus	.02	Music - UI	.05	0.04
			Speech - UI	.02	0.05
Sound Realism	RIR	.48			
	Device	.14			
	Stimulus	< .01	Music - UI	< .01	0.16
			Speech - UI	< .01	0.13

Table 3: Overall ANOVA and post hoc results for each measure in Study 1.

The only significant plausibility findings for the different acoustic conditions were the omnidirectional handclap being less plausible than the Ambisonic RIRs. No significant differences were found when directly comparing the plausibility of 1st- and 3rd-order RIRs, or when directly comparing the influence of impulsive source. The addition of reverberation was found to improve externalisation compared to the dry condition with all RIRs other than the stereo handclap.

Localisation error was found to be significantly lower under the dry condition compared to the majority of RIRs, and when directly comparing the 3rd-order sine and handclap RIRs to analyse the influence of impulsive source, the sine sweep RIR was found to exhibit a higher localisation error. No significant differences were found between 1st- and 3rd-order RIRs for localisation error. Participants also reported higher confidence in their localisations during the dry condition and with the omnidirectional handclap RIR than they did with RIRs of higher spatial resolution. The stereo handclap RIR was also found to result in lower localisation confidence than with the dry, omnidirectional handclap, and 3rd-order handclap RIRs.

Playback device was not found to affect plausibility, but users had higher localisation error, lower localisation confidence, and lower externalisation when using the glasses. We found also that the UI stimulus was the worst-performing of the three test stimuli, with higher localisation error and lower localisation confidence, plausibility, realism, and externalisation than the speech and music stimuli.

5 Study 2: *Sonomancer*, A Real-World AAR Game

The second study (n = 24[11], with an additional 4 participants excluded due to technical glitches or low reliability) focused on the plausibility of the RIRs, and how real-world AAR scenarios affect that.

Participants were taken to two public spaces: a highly reverberant environment (RT = 2.7s), and an outdoor space (RT = 35ms), shown in Figure 1, and asked to play an AAR game over one of the two playback devices. The game, *Sonomancer*, was a localisation game

[11]Final demographics were 13 men, 10 women, and 1 nonbinary, with 2 very unfamiliar, 5 unfamiliar, 6 neutral, 11 familiar, and 4 very familiar with spatial audio.

Fig. 1: Test spaces used in Study 2: outdoor (L), and high reverberance (R).

Measure	Question
Externalisation	Did the game sounds appear to be inside or outside your head?
Plausibility (Brinkmann)	Do you think the [game] sound[s] were recorded in this space? (based on [19])
Plausibility (Definition)	Rate the plausibility of the sound[s] [you heard during the game]
Plausibility (Realism)	Did the [game] sound[s] you heard sound as if they could believably be in this real space?

Table 4: Questions used as part of Study 2. Questions varied slightly between the game and listening test, as shown by the [square brackets].

where the player is tasked with destroying aural monsters as a sonic wizard, or 'Sonomancer'. Participants played game rounds lasting two minutes, and after each round were presented with a short questionnaire, shown in Table 4, asking them to rate the plausibility and externalisation of the sounds they heard in the game. After the game round and questionnaire were completed, they were presented with a blind listening test, based on the process outlined in ITU Recommendation BS.2132-0 [21]. In this listening test, participants were presented with a set of 8-9 audio stimuli, consisting of the speech sample used in Study 1 reproduced using one of the RIRs detailed in Table 1, or a dry condition. Participants were asked to listen to each sample, and rate its plausibility from 0-100 according to one of the three Plausibility questions in Table 4. This listening test then repeated for the other Plausibility questions. Questions in the questionnaire and listening test were presented in a random order, and stimuli were presented in a random arrangement for each listening test. In the first set of listening tests, one of the stimuli was randomly chosen and duplicated to assess rater reliability. Participants who gave significantly different ratings for duplicated stimuli were excluded from the dataset.

This process then repeated until the participant had completed three game rounds and two listening tests, at which point the playback device was swapped. Once the game rounds and listening tests had been completed for both devices, they were taken to the next play space. Test space and playback device were counterbalanced.

Participants localised monsters in the game by turning to face the monster's perceived location, and pulling a trigger on a gamepad controller. If the player localised the monster to within 30°, the monster was successfully 'banished' and the player's score increased by 1. Otherwise, the player missed, and the monster attacked the player, decreasing a virtual health bar. The game featured a variety of sound content, including the synthetic sound of the monster, speech samples recorded by the researchers as a narrator, and a variety of feedback sounds to communicate events such as success or failure, or the end of the game sequence. The game sounds featured reverberation from one of the RIRs chosen at random, and this randomisation was balanced so that all RIRs were presented a roughly equal amount of times across the dataset, as participants would not play enough game rounds to experience all RIRs for a given combination of test space and playback device.

As Study 1 showed minimal differences in plausibility between RIRs, we used two additional plausibility questions to measure the sensation more thoroughly. In contrast to Study 1, no real-world reference for the stimuli used in the game was provided, though as public spaces there were other real-world sounds present. Instead, participants were asked to base their plausibility judgements on their internal reference and expectation for the space's acoustics as this better reflects the user experience of an AAR game or application.

5.1 Results

For Study 2, the three plausibility questions were aggregated together into one measure, as there were no significant differences between how participants answered them in the listening test, and answers were highly correlated ($p < .001$). Plausibility, externalisation, and localisation error were again analysed using a three-way ANOVA (for RIR, playback device, and

Measure		ANOVA p	Significant Pairwise Comparisons	p	Mean Diff.
Plausibility	RIR	< .01	1O-Sine - Dry	.01	4.97
			3O-Sine - Dry	< .01	7.12
			3O-Sine-Echoic - Dry	< .01	5.67
			3O-Sine - Omni-HC	< .01	6.04
			3O-Sine-Echoic - Omni-HC	.03	4.59
			1O-Sine - Stereo-HC	.02	4.66
			3O-Sine - Stereo-HC	< .01	6.81
			3O-Sine-Echoic - Stereo-HC	< .01	5.36
			3O-Sine - 3O-HC	< .01	5.58
			Omni-HC-Echoic - 1O-Sine	< .01	-5.31
			Omni-HC-Echoic - 3O-Sine	< .01	-7.46
			3O-Sine-Echoic - Omni-HC-Echoic	< .01	6
	Device	.38			
	Space	< .01	High Reverb - Outdoor	< .01	5.9
Loc. Error	RIR	< .01	3O-Sine-Echoic - Omni-HC	< .01	5.41
			3O-Sine-Echoic - Stereo-HC	.04	4.29
			3O-Sine-Echoic - 1O-Sine	< .01	5.69
			3O-Sine-Echoic - Omni-HC-Echoic	< .01	5.28
	Device	< .01	Glasses - HD650	< .01	2.30
	Space	0.21			
Extern.	RIR	.94			
	Device	.70			
	Space	.38			

Table 5: Overall ANOVA and post hoc results for each measure in Study 2.

test space), and *post hoc* analysis was conducted using Tukey HSD tests. Results are shown in Table 5.

For plausibility ratings given in the listening test, a significant difference between the 3rd-order sine sweep and handclap RIRs was found, with the sine sweep impulse source being rated more plausible than the handclap. Analysis of the factor interaction between RIR and test space revealed this only applied in the outdoor space ($p < .01$). No difference in plausibility was found when directly comparing the 1st-order sine sweep RIR with its 3rd-order counterpart. Echoicity was also not found to impact plausibility, with no significant difference between plausibility ratings given for either pair of echoic and anechoic RIRs.

As part of the game (where participants did not experience all RIRs, and so comparisons are between-subjects), the only finding for localisation error was the echoic 3rd-order sine sweep RIR having a higher error than some other RIRs, such as the 1st-order sine sweep. When directly comparing 1st- and 3rd-order RIRs, impulsive sources, or echoic/anechoic pairs, localisation error was not found to be affected. No influence of RIR was found on externalisation ratings given in the game.

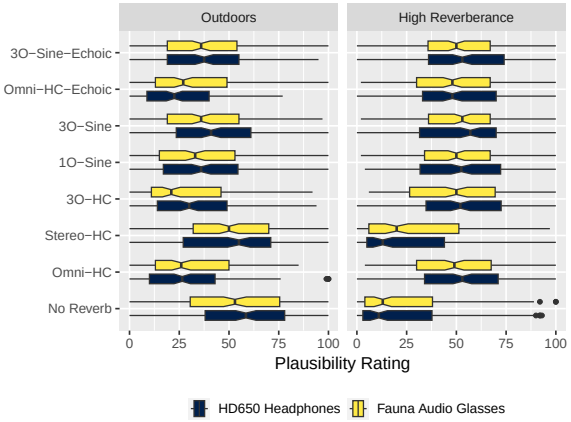

Fig. 2: Boxplot of plausibility ratings given as part of the listening test in Study 2, separated by space, RIR, and playback device.

Plausibility data were also gathered as part of the post-game questionnaire. Comparing the plausibility ratings given for an RIR after gameplay and as part of the listening test, participants rated sounds as being significantly more plausible during the game than as part of the listening test ($p < .01$, with mean rating of 65.4 in game and 42.25 in listening test), as shown in Figure 3.

In the listening test, plausibility was not found to be affected by playback device, but when playing the game, sounds were rated as less plausible for glasses than headphones (mean rating 66.57 for HD650, 63.94 for glasses, $p = .3$). Participants also had a higher localisation error in the game using glasses, though externalisation was not found to be affected. Sounds presented in the high reverberance environment were also found to be slightly more plausible than those presented outdoors. No other effect of test space was found.

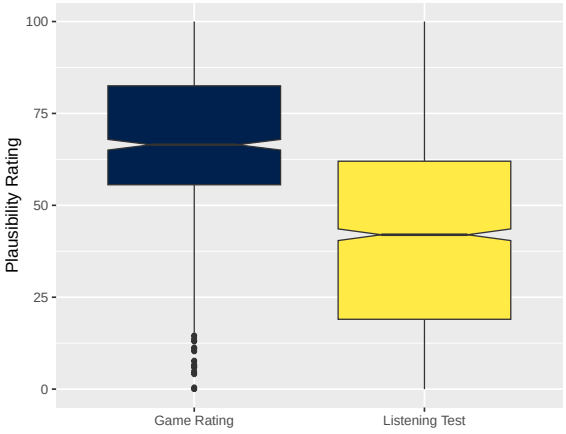

Fig. 3: Plausibility ratings given for RIRs as part of the game experience and then as part of the listening test.

6 Interview Results

Post-study interviews were carried out for both studies, inviting participants to discuss their experiences of the two playback devices and the different acoustic conditions. In both interviews, participants were asked which of the two playback devices they would prefer to use in a future AAR application, with guided museum tours, auditory navigation, and AAR games being provided as example scenarios. In both, participants preferred using the audio glasses, praising their form factor and some being pleasantly surprised by their audio quality. In Study 1, 10 participants preferred glasses compared to 7 for headphones, and in Study 2 13 preferred glasses, and 8 preferred headphones. Participants often indicated their preference was contextual (3 in Study 1 and 9 in Study 2), particularly based on background noise, noting a degraded experience using audio glasses in the noisier high reverberance space in Study 2.

Participants who preferred the headphones often cited superior sound quality, ease of localisation, and their familiarity as being important factors. Many participants also indicated that the glasses helped them feel more 'connected' to their real-world auditory surroundings (3 in Study 1 and 10 in Study 2), or that headphones isolated them from their surroundings (5 in Study 1 and 9 in Study 2). This acoustic transparency was noted as being a positive by 11 participants and a negative aspect by 5 participants. While acoustic transparency is core to facilitating AAR, this was not mentioned directly to participants to avoid biasing them.

"I would have thought that before the study, the headphones would have immersed me deeper, but it was the glasses...the glasses sort of put me into the game, whereas the headphones sort of took me out." (P8, Study 2)

Participants noted that the inclusion of reverberation improved the plausibility of virtual sounds, and that it affected their ability to localise sounds in both studies – 13 positively and 9 negatively. When asked about differences between conditions, some mentioned certain conditions being clearly worse or less real than others, but participants often noted that any perceived differences were minimal, with one participant even noting they did not realise different conditions existed.

"I really struggled to tell the difference between them, to be honest. I think I could tell what seemed like a...kinda...stronger reverb, I guess? But the subtleties of it were lost on me." (P2, Study 2)

7 Discussion and Conclusions

A number of conclusions can be drawn from both studies as to the influence of spatial resolution and spectral bandwidth on virtual sound perception. Both studies show minimal differences between plausibility ratings for the different RIRs, suggesting that overall, for the tasks and situations chosen for our studies, our users are not particularly sensitive to differences in acoustic reproductions. Ambisonic RIRs resulted in more plausible presentation, but as neither study showed 3rd-order RIRs to be significantly more plausible than 1st-order counterparts, we would suggest developers consider 1st-order RIRs for acoustic reproductions when modelling direct sound separately. Study 1 further suggested minimal sensitivity to differences in reverberation through its externalisation findings, where the inclusion of reverberation improved externalisation but there were no differences between reverberant conditions.

Study 2 provides some practical guidance on choosing RIRs for plausible playback. With the 3O-Sine RIR rated as more plausible than the 3O-HC, it suggests that higher spectral bandwidth can make a tangible improvement to plausibility, and that AAR developers should consider prioritising this. Study 2 also shows that echoicity (at the levels we tested), does not influence plausibility, and therefore that AAR developers can employ similarly echoic sounds, be they from

field or foley recordings or existing libraries. Finally, Study 2 shows that application context significantly influences plausibility, finding in-game sounds to be more plausible than those in the listening test. While these game ratings were given retrospectively and for different sound content, it suggests that developers can afford to consider simpler, more computationally efficient reproductions, and that findings from critical listening tests may not represent real-world performance.

Notably, the Stereo-HC condition performed poorly in both studies. We had expected Stereo-HC to perform partway between the Omni-HC and 3O-HC conditions to reflect its slightly higher spatial resolution, but it actually performed closer to the dry conditions in both studies. As the Stereo-HC condition was rendered by taking signals from opposing capsules on the Zylia microphone, it is possible that these capsules have a different pickup pattern or are pre-processed by the microphone's onboard interface when recording, and are therefore not representing a stereo microphone in the way we had intended. Future work could explore how other stereo RIRs influence plausibility, as they may still prove a midpoint between the accessibility of an omni RIR and the perceptual benefits of Ambisonics.

Another limitation to acknowledge is that both studies featured egocentric sounds from a static listening position. In an exocentric scenario where users can freely move around a space, the greater spatial accuracy of higher order RIRs may be more influential, and this could be another interesting avenue for future work.

As one of the first perceptual explorations of audio glasses, our results provide some interesting findings on their user experience. Firstly, neither study found an influence of playback device on plausibility. Study 2 suggested audio glasses may be slightly less plausible in an application scenario, though the difference was not very large (3%). Both studies' findings also suggest that glasses are perceptually worse than headphones, with a higher localisation error in both studies, and lower levels of externalisation in Study 1. That said, the interview data from both studies suggest that glasses are a promising platform for AAR and one that users are interested in. It is important also to note that as we did not test multiple sets of headphones and audio glasses, these results could be specific to these models. There is also a large price difference between the two (with the HD650 headphones retailing for two to three times more than the FAUNA glasses), which could further exaggerate the perceptual differences we found. Overall, our work shows that reproducing an environment's acoustics provides a tangible step towards the sensation of a virtual sound being located seamlessly and believably in a user's surroundings, key for AAR applications. This can be done in a simpler way that reduces processing power and battery consumption, and can be presented over both novel and traditional hardware, be that for audio augmented reality, extended reality experiences, games, or beyond.

8 Acknowledgement

This work was supported by the SONICOM project (www.sonicom.eu), funded by the Horizon 2020 programme under grant agreement 101017743.

References

[1] Yang, J., Barde, A., and Billinghurst, M., "Audio Augmented Reality: A Systematic Review of Technologies, Applications, and Future Research Directions," *Journal of the Audio Engineering Society*, 70(10), pp. 788–809, 2022, ISSN 15494950, doi:10.17743/jaes.2022.0048.

[2] Lindau, A. and Weinzierl, S., "Assessing the Plausibility of Virtual Acoustic Environments," *Acta Acustica united with Acustica*, 98(5), pp. 804–810, 2012, ISSN 16101928, doi:10.3813/AAA.918562.

[3] McGill, M., Brewster, S., McGookin, D., and Wilson, G., "Acoustic Transparency and the Changing Soundscape of Auditory Mixed Reality," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pp. 1–16, Association for Computing Machinery, New York, NY, USA, 2020, ISBN 978-1-4503-6708-0, doi:10.1145/3313831.3376702.

[4] Mariette, N., "Human Factors Research in Audio Augmented Reality," in W. Huang, L. Alem, and M. A. Livingston, editors, *Human Factors in Augmented Reality Environments*, pp. 11–32, Springer, New York, NY, 2013, ISBN 978-1-4614-4205-9, doi:10.1007/978-1-4614-4205-9_2.

[5] Roginska, A., "Binaural Audio Through Headphones," in A. Roginska and P. Geluso, editors, *Immersive Sound*, pp. 88–123, Routledge, 1 edition, 2017, ISBN 978-1-315-70752-5, doi:10.4324/9781315707525-5.

[6] Lyons, K., Gandy, M., and Starner, T., "Guided by Voices: An Audio Augmented Reality System," in *Proceedings of the 6th International Conference on Auditory Display*, 2000.

[7] Chatzidimitris, T., Gavalas, D., and Michael, D., "SoundPacman: Audio Augmented Reality in Location-Based Games," in *2016 18th Mediterranean Electrotechnical Conference (MELECON)*, pp. 1–6, IEEE, 2016, ISSN 2158-8481, doi:10.1109/MELCON.2016.7495414.

[8] Rovithis, E., Moustakas, N., Floros, A., and Vogklis, K., "Audio Legends: Investigating Sonic Interaction in an Augmented Reality Audio Game," *Multimodal Technologies and Interaction*, 3(4), p. 73, 2019, ISSN 2414-4088, doi:10.3390/mti3040073.

[9] Paterson, N., Naliuka, K., Jensen, S. K., Carrigy, T., Haahr, M., and Conway, F., "Spatial Audio and Reverberation in an Augmented Reality Game Sound Design," in *Audio Engineering Society Conference: 40th International Conference: Spatial Audio: Sense the Sound of Space*, p. 9, Audio Engineering Society, 2010.

[10] "ISO 3382-1: 2009. Measurement of Room Acoustic Parameters," 2009.

[11] Best, V., Baumgartner, R., Lavandier, M., Majdak, P., and Kopčo, N., "Sound Externalization: A Review of Recent Research," *Trends in Hearing*, 24, 2020, ISSN 2331-2165, 2331-2165, doi:10.1177/2331216520948390.

[12] Blauert, J., *Spatial Hearing: The Psychophysics of Human Sound Localization*, MIT press, 1997.

[13] Ahrens, J. and Andersson, C., "Perceptual Evaluation of Headphone Auralization of Rooms Captured with Spherical Microphone Arrays with Respect to Spaciousness and Timbre," *The Journal of the Acoustical Society of America*, 145(4), pp. 2783–2794, 2019, ISSN 0001-4966, 1520-8524, doi:10.1121/1.5096164.

[14] Frank, M., Zotter, F., and Sontacchi, A., "Producing 3D Audio in Ambisonics," in *Audio Engineering Society Conference: 57th International Conference: The Future of Audio Entertainment Technology–Cinema, Television and the Internet*, Audio Engineering Society, 2015.

[15] Enge, K., Frank, M., and Holdrich, R., "Listening Experiment on the Plausibility of Acoustic Modeling in Virtual Reality," *Fortschritte der Akustik–DAGA*, pp. 13–16, 2020.

[16] Engel, I., Henry, C., Amengual Garí, S. V., Robinson, P. W., and Picinali, L., "Perceptual Implications of Different Ambisonics-based Methods for Binaural Reverberation," *The Journal of the Acoustical Society of America*, 149(2), pp. 895–910, 2021, ISSN 0001-4966, doi:10.1121/10.0003437.

[17] Cuevas-Rodríguez, M., Picinali, L., González-Toledo, D., Garre, C., de la Rubia-Cuestas, E., Molina-Tanco, L., and Reyes-Lecuona, A., "3D Tune-In Toolkit: An Open-Source Library for Real-Time Binaural Spatialisation," *PLOS ONE*, 14(3), 2019, ISSN 1932-6203, doi:10.1371/journal.pone.0211899.

[18] Engel, I., Daugintis, R., Vicente, T., Hogg, A. O. T., Pauwels, J., Tournier, A. J., and Picinali, L., "The SONICOM HRTF Dataset," *Journal of the Audio Engineering Society*, 71(5), pp. 241–253, 2023, ISSN 15494950, doi:10.17743/jaes.2022.0066.

[19] Brinkmann, F., Aspöck, L., Ackermann, D., Lepa, S., Vorländer, M., and Weinzierl, S., "A Round Robin on Room Acoustical Simulation and Auralization," *The Journal of the Acoustical Society of America*, 145(4), pp. 2746–2760, 2019, ISSN 0001-4966, doi:10.1121/1.5096178.

[20] Poirier-Quinot, D., Lawless, M. S., Stitt, P., and Katz, B. F., "HRTF Performance Evaluation: Methodology and Metrics for Localisation Accuracy and Learning Assessment," in *Advances in Fundamental and Applied Research on Spatial Audio*, IntechOpen, 2022, ISBN 978-1-83969-005-1 978-1-83969-006-8, doi:10.5772/intechopen.104931.

[21] "RECOMMENDATION ITU-R BS.2132-0 – Method for the Subjective Quality Assessment of Audible Differences of Sound Systems Using Multiple Stimuli without a given Reference," 2019.

Full dataset DOI:10.5281/zenodo.10605358